Personality Trait Classification of Essays with the Application of Feature Reduction

Edward P. Tighe, Jennifer C. Ureta, Bernard Andrei L. Pollo,
Charibeth K. Cheng, and **Remedios de Dios Bulos**
De La Salle Univeristy, Manila, Philippines
{edward_p_tighe, jennifer_ureta, bernard_pollo}@dlsu.edu.ph,
chari.cheng@delasalle.ph, remedios.bulos@dlsu.edu.ph

Abstract

Determining an individual's personality traits is an important concept in Psychology. Although traits are normally assessed through self-report tests, an alternative method would be to computationally analyze an individual's linguistic markers. Studies in personality trait classification show promising results and look to continuously improve the field by either using new features or by collecting new data from social media; however, a key concept that is not always considered is the use of feature reduction techniques. This research aims to perform feature reduction techniques on linguistic features from essays and classify the author's personality traits based on the reduced feature set. The classifiers are evaluated by comparing against a baseline classifier trained with all extracted features. The feature reduction techniques used are Information Gain and Principal Component Analysis. The results show that feature reduction techniques are able to increase classification measures, but not by significant values. Reduced datasets are exceptionally beneficial in reducing the amount of data needed allowing classifiers to perform faster while still maintaining classification measures.

1 Introduction

Personality Psychology, or simply Personality, is "the scientific study of psychological forces that make people uniquely themselves" [Friedman and Schustack, 2014, p.1]. These forces consist of organized and relatively enduring traits and mechanisms that influence one's interactions with the intrapsychic, physical, and social environments [Larsen and Buss, 2008, p.4].

One of the most well-researched theories describing personality trait variation would be the Five Factor model, also known as the Big Five [Norman, 1963; Goldberg, 1981]. The Big Five is an organization of personality facets that are subsets of five broad traits: *Extraversion*, *Agreeableness*, *Conscientiousness*, *Neuroticism*, and *Openness to Experience*. John et al. [2008, p.116] mentions that these five traits "were derived from analyzing terms people use to describe themselves and others." The Big Five have been used in studies observing individuals in environments such as at work [Richardson *et al.*, 2009] and in academics [Komarraju *et al.*, 2009]. Research using both natural language adjectives and theoretically based personality instruments supports the comprehensiveness of the model and its applicability across observers and cultures [McCrae and John, 1992].

Personality traits are traditionally measured through the use of questionnaires such as the Big Five Inventory (BFI) [John *et al.*, 1991]; however, an alternative approach would be to analyze an individual's linguistic markers. An individual's choice of words eventually becomes consistent over time and context and can be used as an individual difference measure [Pennebaker *et al.*, 2003]. A study by Pennebaker and King [1999] showed multiple correlations between linguistic markers and the Big Five such as how Neuroticism is positively correlated with the use of negative emotion words and negatively with positive emotion words. Goldberg [1981, p.142] mentions that "the more important an individual difference [is] in human transactions; the more likely languages will have a term for it."

Correlations between linguistic markers and personality traits have paved the way for research in the area of automatic personality classification. One of the earliest studies [Mairesse *et al.*, 2007] focused on classifying personality traits based on text. They extracted linguistic features from essays using a text analysis tool and a psycholinguistic database. Their findings were modest, but still showed that computationally modeling the Big Five was possible. Subsequent studies were able to present promising methods in improving upon the findings of Mairesse et al. [2007] by introducing new linguistic features [Mohammad and Kiritchenko, 2013; Poria *et al.*, 2013a]. Other studies [Golbeck *et al.*, 2011; Schwartz *et al.*, 2013; Park *et al.*, 2014; Peng *et al.*, 2015] focus on different data sources by taking advantage of social media, and collecting or using data from users of these growing platforms.

Although there have been advancements in the field of personality trait classification, there are still gaps in determining which linguistic features are most significant for the classification process. This research investigates the use of two feature reduction techniques in order to improve the computation involved. The data source for this research is the Pennebaker and King [1999] dataset of essays. Features are extracted us-

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 22-28, New York City, USA, July 10, 2016.

ing LIWC and are analyzed to see if they can be reduced to a smaller set while still being able to aid in classifying the Big Five. The classifiers are then built using the reduced set of features and are compared against classifiers using the complete set of features. This research aims to show that the use of feature reduction techniques are beneficial to future work in the field.

The remainder of this paper is organized as follows: Section 2 reviews studies on personality trait classification that work with the Pennebaker and King [1999] essay dataset. Section 3 discusses the characteristics of the data source used in this research. Section 4 explains how features are extracted from the data source. Section 5 explains the background of the feature reduction techniques used. Section 6 presents how each classifier was built. Section 7 discusses the overall performance of the classifiers and the effects of feature reduction. Finally, Section 8 concludes the paper and explains recommendations for future work.

2 Related Works

Personality Trait Classification based on linguistic markers is a growing field. Although other studies [Golbeck *et al.*, 2011; Schwartz *et al.*, 2013; Park *et al.*, 2014; Peng *et al.*, 2015] use big collections of data from social media, this paper limits its review to studies that use the Pennebaker and King [1999] essay dataset for the sake of having a common ground.

One of the earliest studies regarding automatic personality classification is that of Mairesse et al. [2007]. Their methods of extracting features relied on LIWC[1] [Pennebaker *et al.*, 2001] and the MRC Psycholinguistic Database [Coltheart, 1981]. LIWC produced a total of 88 features and is further discussed in Section 4. They also used 14 psycholinguistic features[2] from the MRC Psycholinguistic Database, a machine usable dictionary. They then trained classifiers based on different combinations of the set of features and had promising results. Openness to Experience was the easiest to identify among the Big Five having an accuracy of 62.5% using only LIWC features. They also showed how features from LIWC out performed those from MRC; however, both showed promising correlations to the Big Five. Their results were modest, but were significant enough to show that computationally modeling personality traits was possible.

One recent study [Mohammad and Kiritchenko, 2013] made use of fine affect or emotion category features as alternatives for personality trait classification. They were able to extract an extensive amount of emotion features with the use of the NCR Hashtag Emotion Lexicion [Mohammad and Turney, 2010]. This lexicon is able to produce either 8 basic emotions or 585 fine emotion features. They also made use of the Specificity Lexicon and Osgood Dimensions Lexicon [Turney and Littman, 2003]. The first lexicon calculated the average information content of an essay while the later was able to extract the average evaluativeness, activity, and potency scores of words. Finally, they made use of the LIWC features and frequencies of unigrams of the essays. They experimented by combining different features sets and ran them through Support Vector Machine classifiers. The classifier that performed best was built using the LIWC and the 585 fine emotion features. Their results showed minimal improvement over the results of Mairesse et al. [2007], but revealed that emotion category features contain information regarding an individual‘s personality and can be considered useful for future studies.

Another study [Poria *et al.*, 2013a] introduced a novelty approach of using of common sense knowledge. They utilize ConceptNet [Havasi *et al.*, 2007] and EmoSenticNet [Poria *et al.*, 2013b] to extract sentiment polarity scores and affective labels from the essays. They also extract linguistic features from LIWC and MRC. They train Support Vector Machine classifiers and compare against Mairesse et al [Mairesse *et al.*, 2007] and Mohammad and Kiritchenko [Mohammad and Kiritchenko, 2013]. Their results show significant improvements demonstrating that the sentiment polarity and affective labels contained relevant information in classifying personality traits.

With the discovery of more and more features with information pertaining to an individual‘s personality, the issue of irrelevant features and overfitting arises. Each of the previously reviewed studies presents an opportunity to investigate the use of feature reduction due to the high volume of features presented. This research explores the application of feature reduction on LIWC features and aims to showcase the benefits of using these techniques.

3 Data Source

The data used in this research was gathered and used in a study by Pennebaker and King [1999]. The actual file was retrieved from myPersonality[3]. It consists of a total of 2,468 essays or daily writing submissions from 34 psychology students. There are a total of 29 women and 5 men whose ages ranged from 18 to 67 with a mean of 26.4 and a standard deviation of 11.1.

The writing submissions were in the form of a course requirement or assignment but were not graded. For each assignment, students were expected to write a minimum of 20 minutes per day about a specific topic. The data was collected during a 2-week summer course between 1993 to 1996. Each student completed their daily writing for 10 consecutive days.

Students' personality scores were assessed by answering the Big Five Inventory (BFI) [John *et al.*, 1991]. The BFI is a 44-item self-report questionnaire that provides a score for each of the five personality traits. Each item consists of short phrases and is rated using a 5-point scale that ranges from 1 (disagree strongly) to 5 (agree strongly).

An instance in the data source consists of a filename or ID, the actual essay, and five classification labels of the Big Five personality traits. Labels were originally in the form of either yes (‘y’) or no (‘n’) to indicate scoring high or low for a given trait; however, this research changed the labels to ‘y’ to 1 and ‘n’ to 0 according to the preference of the researchers.

[1] It was assumed by the researchers that Mairesse et al. [2007] used the 2001 version of LIWC as to how it was cited in their paper

[2] However, the total number of features is listed as 26 [Wilson, 1988]

[3] www.mypersonality.org

4 Feature Extraction

In order to extract information from raw text, LIWC[4] was utilized. LIWC stands for Linguistic Inquiry and Word Count and was developed by Pennebaker et al. [2007]. It is a text analysis tool that provides an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' written samples.

The tool analyses text files sequentially, one target word at a time by searching through it dictionary file. If the target word matches the dictionary word, the appropriate word category scale is incremented. Pennebaker et al. [2007] explains that there are a total of 80 output features consisting of 4 general descriptor categories (e.g., total word count, words per sentence), 22 standard linguistic dimensions (e.g., frequency of pronouns, articles), 32 word categories tapping psychological constructs (e.g., affect, cognition), 7 personal concern categories (e.g., work, home), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (e.g., periods, commas). Values of all features except word count and words per sentence reflect percentage of total words.

Based on the methods of Mairesse et al. [2007], the essays from the data source were fed through LIWC. The output features were used to create a dataset for each of the Big Five. These datasets contain all 80 LIWC features and one of the five personality trait as the classification label. It is important to note that replication of the methods used by Mairesse et al. [2007] was chosen by the researchers as the better alternative to direct comparison of results. This was due to the difference in the number of output features from the version of LIWC reported in their paper. Replication of methods would allow for a better baseline when comparing against a reduced set of features.

5 Feature Reduction

LIWC provides a vast amount of information that it would be important to analyze whether or not classification of the Big Five can be improved by reducing the set of features. The presence of non-relevant features can influence a classifier to produce smaller error by fitting the model according to the training data. Removing such features can increase the predictive power of a classifier by focusing only on certain features. A defined model may be able to classify unseen data better and is desirable for real world scenarios. Feature reduction becomes an important concept to consider when trying to improve classification. The techniques that are performed are Information Gain and Principal Component Analysis.

The following subsections discuss how these techniques work and present their respective output. Both techniques were applied on the datasets using Waikato Environment for Knowledge Analysis or Weka, a tool of machine learning algorithms and data preprocessing [Hall *et al.*, 2009].

5.1 Information Gain

Information Gain is a measure of how effective a given feature is in classifying data [Mitchell, 1997]. It becomes essential to this research to evaluate each of the 80 LIWC features and determine which provide significance in classifying the Big Five.

One important concept in computing for the Information Gain is Entropy or being able to characterizes the impurity of an arbitrary collection of examples. Entropy is defined as

$$E(F) = -\sum_{v=1}^{n} p_v \log_2 p_v$$

where F is a feature or classification containing a number n of different discrete values and where p_v is the proportion of F belonging to value v. When values are continuous in nature, the values are discretized by splitting at a point that provides the maximum information gain. Therefore, the information gain of a feature F relative to a classification C can be defined as

$$IG(C, F) = E(C) - \sum_{v \in Values(F)} \frac{|C_v|}{|C|} E(C_v)$$

where $Values(F)$ is the set of all possible values for feature A, and C_v is the subset of C for which feature F has values v. Basically, information gain is the entropy of class C reduced by the weighted average entropy of each subset S_v.

The Information Gain of all 80 features were computed for each of the Big Five datasets. For each of the datasets, only features with non-zero information gain were selected. Table 1 shows the remaining features per personality trait along with their respective information gain.

5.2 Principal Component Analysis

Principal Components Analysis (PCA) is used to identify patterns, and highlight the similarities and differences in data [Smith, 2002]. It is particularly useful when dealing with data with a high number of features as it is able to reduce the number of these features without losing much information. Concepts of covariance, matrix operations, eigenvalues, and eigenvectors are all used to compute for the principal components.

Smith [2002] explains that in order to perform PCA, the first step would be to calculate all the features' covariance matrix CM which can be defined for a set of data with m features as

$$CM^{m \times m} = \begin{pmatrix} cov(F_1, F_1) & \cdots & cov(F_1, F_m) \\ \vdots & \ddots & \vdots \\ cov(F_m, F_1) & \cdots & cov(F_m, F_m) \end{pmatrix}$$

where $CM^{m \times m}$ is a matrix with m rows and m columns composed the covariance between features Fx where x ranges from 1 to m. The second step is to calculate the CM's eigenvectors and their respective eigenvalue. Once found, the eigenvalues are ranked from highest to lowest and are removed along with their paired eigenvector according to a set threshold. The remaining eigenvectors are then inserted into a Feature Vector FV from highest to lowest eigenvalue. The final dataset values $FinalData$ is defined as

$$FinalData = FV^{\mathsf{T}} \times AdjustedValues^{\mathsf{T}}$$

[4] Developed in 2007 and is a different version than the previously mentioned LIWC in Section 2

Table 1: The remaining features for each of the Big Five after removing LIWC features with zero information gain

Extraversion		***Conscientiousness***			
Gain	*Feature*	*Gain*	*Feature*	*Gain*	*Feature*
0.00696	Articles	0.00996	Swear words	0.00688	Apostrophes
0.00643	Personal Pronouns	0.00975	Anger	0.00684	Function Words
0.00637	Sexuality	0.00825	Negative Emotion	0.00659	Prepositions
0.00589	Conjunctions	0.00731	Dictionary Words	0.00659	Exclamation Marks
Openness to Experience					
Gain	*Feature*	*Gain*	*Feature*	*Gain*	*Feature*
0.01583	Dictionary Words	0.00836	Friends	0.00728	Personal Pronouns
0.01496	Work	0.00808	All Punctuation	0.00712	Cognitive Processes
0.01300	First Person Singular Pronoun	0.00807	Motion	0.00661	Function Words
0.01216	Second Person Pronoun	0.00802	Religion	0.00655	Words $>$ 6 Letters
0.01046	Home	0.00763	Parentheses	0.00644	Question Marks
0.00978	Time	0.00760	Articles	0.00623	Sexuality
0.00937	Relativity	0.00736	Commas	0.00552	Quotation Marks
0.00844	Swear Words	0.00735	Regular Verbs	0.00534	Anger
Agreeableness		***Neuroticism***			
Gain	*Feature*	*Gain*	*Feature*	*Gain*	*Feature*
0.01103	Anger	0.01948	Negative Emotion	0.00647	Dictionary Words
0.00987	Swear Words	0.00911	Sadness	0.00635	Total Pronouns
0.00708	Negative Emotion	0.00895	First Person Singular Pronoun	0.00595	Negations
0.00680	Family	0.00750	Anxiety	0.00555	Leisure
0.00647	Dictionary Words	0.00744	Personal Pronoun		

where the transposed FV is multipled with a transposed matrix $AdjustedValues$ containing the original datasets's values adjusted by each feature's mean.

After performing Weka's implementation of PCA on all 80 LIWC features, a total of 56 eigenvectors were found and used to create a new dataset for each of the Big Five.

6 Classification

A 10-fold cross validation was performed on each of the 15 datasets (5 using all features, 5 using information gain, and 5 using PCA) in order to evaluate their overall effectiveness. This research recorded the accuracy, precision, and F-measure of all classifiers and the amount of reduction in terms of a dataset's feature size.

Each dataset was fed through three learning algorithms in Weka and compared against a baseline classifier (*ZeroR*) that returned the majority class. The algorithms used are two implementations of Support Vector Machine (*libSVM* and *SMO*), and Linear Logistic Regression (*SimpleLogistic*). Other algorithms such as k-Nearest Neighbour (*IBk*, where k equaled *1* and *5*), C4.5 Decision Tree (*J48*), Naive Bayes (*NaiveBayes*), and Random Forest (*RandomForest*) were also investigate; however, these classifiers were discarded due to poor performance. Default parameter settings were used for each of the learning algorithms.

7 Discussion

An overview of the results of classification, as seen in Table 2, shows that Openness to Experience is the easiest trait to identify regardless of feature reduction techniques. The remaining traits, ranked from easiest to hardest to identify, are Neuroticism, Agreeableness, Conscientiousness, and Extraversion. The ranking corresponds to the Information Gain of each LIWC feature per Big Five as seen in Table 1. Openness to Experience had the most amount of features remaining after removing those with zero information gain. On the other hand, Extraversion had the least remaining features. Each of the remaining features can also serve as a descriptor of how word choice of an individual is related to their personality traits. The use of *Negative Emotions* is relevant in determining one's Conscientiousness, Agreeableness, and Neuroticism. Similarly, *Swear Words* is relevant to Conscientiousness, Openness to Experience and Agreeableness. Although, it is important to note that that higher usage rate of a linguistic feature does not equates to a high personality trait score. Features with higher Information Gain simply indicate that a feature is more effective in separating and classifying data.

The classifiers using feature reduced datasets were generally able to increase classification measures, but not by significant values. This suggests that the LIWC features do not have any more information to provide in classifying the Big Five, at least when considering only a feature set of only LIWC features. A better comparison of the classifiers built using all 80 LIWC features and the feature reduced datasets is shown in Table 3. This indicates that classifiers using feature reduced datasets were able to slightly edge out classifiers using all features in four of the five personality traits. Agreeableness was the only trait where both best classifiers performed similarly.

This research also noted the minimal increase in classi-

Table 2: The performance of each classifier according to their respective dataset

Personality Trait	Features Used	Number of Feature	Size Reduction	Classifier	Accuracy	Precision	F-measure
Agreeableness	All Features	80	0.00%	ZeroR	53.08%	0.282	0.368
				LibSVM	51.78%	0.502*	0.475*
				SMO	56.20%	0.559*	0.549*
				SimpleLogistic	57.42%*	0.572*	0.566*
	Information Gain	5	93.75%	LibSVM	55.47%	0.551	0.536**
				SMO	55.88%	0.578	0.487
				SimpleLogistic	57.54%	0.575	0.557
	PCA	56	30.00%	LibSVM	57.05%**	0.568**	0.563**
				SMO	55.15%	0.547	0.535
				SimpleLogistic	55.55%	0.552	0.547
Conscientiousness	All Features	80	00.00%	ZeroR	50.81%	0.258	0.342
				LibSVM	51.99%	0.520*	0.520*
				SMO	54.82%	0.548*	0.545*
				SimpleLogistic	54.91%*	0.549*	0.548*
	Information Gain	8	90.00%	LibSVM	53.73%	0.537	0.537
				SMO	55.55%	0.560	0.541
				SimpleLogistic	55.27%	0.553	0.550
	PCA	56	30.00%	LibSVM	56.04%**	0.560**	0.560**
				SMO	55.80%	0.558	0.554
				SimpleLogistic	55.39%	0.554	0.552
Extraversion	All Features	80	00.00%	ZeroR	51.74%	0.268	0.353
				LibSVM	51.22%	0.507*	0.492*
				SMO	53.85%	0.537*	0.533*
				SimpleLogistic	53.49%	0.533*	0.530*
	Information Gain	4	95.00%	LibSVM	52.82%	0.527	0.518
				SMO	54.74%	0.547	0.532
				SimpleLogistic	54.70%	0.546	0.541
	PCA	56	30.00%	LibSVM	55.75%**	0.557**	0.556**
				SMO	53.65%	0.535	0.531
				SimpleLogistic	53.00%	0.528	0.527
Neuroticism	All Features	80	00.00%	ZeroR	50.04%	0.250	0.334
				LibSVM	51.09%	0.511*	0.509*
				SMO	57.05%	0.571*	0.570*
				SimpleLogistic	57.46%*	0.575*	0.575*
	Information Gain	9	88.75%	LibSVM	55.79%**	0.558**	0.558**
				SMO	57.13%	0.572	0.570
				SimpleLogistic	57.45%	0.575	0.574
	PCA	56	30.00%	LibSVM	58.31%**	0.583**	0.583**
				SMO	56.69%	0.567	0.567
				SimpleLogistic	57.17%	0.572	0.572
Openness to Experience	All Features	80	00.00%	ZeroR	51.54%	0.266	0.351
				LibSVM	54.05%	0.540*	0.524*
				SMO	61.26%*	0.613*	0.613*
				SimpleLogistic	59.52%*	0.595*	0.595*
	Information Gain	24	70.00%	LibSVM	56.93%	0.569	0.567
				SMO	61.83%	0.618	0.618
				SimpleLogistic	61.06%	0.610	0.610
	PCA	56	30.00%	LibSVM	59.92%	0.599	0.598
				SMO	61.95%	0.619	0.619
				SimpleLogistic	61.83%	0.618	0.618

Paired t-test was performed in Weka where significance was set at 0.05

*Significantly better than the baseline (*ZeroR*); **Significantly better than its respective classifier trained using all features

Table 3: Comparison of best performing classifiers using all features and using feature-reduced datasets

	Using All Features				Using Feature-Reduced Datasets			
Big Five	**Classifier**	**Accuracy**	**Precision**	**F-measure**	**Classifier**	**Accuracy**	**Precision**	**F-measure**
Agreeableness	SimpleLogistic	57.42%	0.572	0.566	SimpleLogistic[A]	57.54%	0.575	0.557
Conscientiousness	SimpleLogistic	54.91%	0.549	0.548	LibSVM[B]	56.04%	0.560	0.560
Extraversion	SMO	53.85%	0.537	0.533	LibSVM[B]	55.75%	0.557	0.556
Neuroticism	SimpleLogistic	57.46%	0.575	0.575	LibSVM[B]	58.31%	0.583	0.583
Openness to Experience	SMO	61.26%	0.613	0.613	SMO[B]	61.95%	0.619	0.619

Classifiers with [A] were trained using Information Gain reduced feature sets and [B] represents classifiers trained using PCA reduced feature sets

fication measures from using all features to using reduced features when dealing with Agreeableness. This can be attributed to how the classifier *SimpleLogistic* works. It is important to note that *SimpleLogistic* includes its own implementation of feature reduction [Landwehr *et al.*, 2005]. Interestingly, the only common attributes in comparison to those selected using Information Gain were *Anger* and *Family*. The remaining attributes selected by *SimpleLogistic* were *Words greater than 6 letters*, *Common Adverbs*, *Negations*, *Anxiety*, *Motion*, *Exclamation Marks*, and *Dashes*. The classifier also selected features for other personality traits that were not present in the set of remaining features after performing Information Gain.

Despite the results that feature reduction techniques were not able to significantly increase classification measures, it is important to note that the amount of reduction made in the size of the set of features is significant as seen in column *Size Reduction* of Table 2. Datasets using Information Gain had highly significant size reductions ranging from *70%* to *95%* while still able to perform up to par with those using all 80 LIWC feature. On the other hand, PCA was able to significantly reduce the set of features by *30%* while still covering *95%* of the feature set's variance. The PCA reduced dataset was also able to significantly improve the classifier *LibSVM* for all personality traits which is evident when looking at the measures. Agreeableness was able to improve from an F-measure of 0.475 to 0.563 resulting in a 0.088 increase. On the other hand, Conscientiousness had improved from an F-measure of 0.52 to 0.56 resulting in the the lowest increase of 0.04. As a whole, feature reduction techniques were able to improve the classification of personality traits by both slightly increasing classification measures and heavily reducing the size of the datasets.

8 Conclusion and Recommendations

This research was able to demonstrate that feature reduction techniques like Information Gain and Principal Component Analysis are beneficial in classifying an individual's personality traits based on text data. Applying these techniques reduced the size of the original data while slightly improving the classifiers' level of performance. A reduced-dataset leads to a more defined model which can better handle unseen data. This research was also able to highlight LIWC features that contain the most Information Gain about an individual's traits. This knowledge can be useful outside of computational classification by providing additional linguistic descriptors of individual's with certain personality traits.

This research also recommends two areas for improvements regarding future work in the field of text-based personality trait classification. The first would concern the data source containing binomially labeled traits and thereby categorizing an individual into one or the other. This representation does not capture the dimensional nature of a trait and would be better represented as either the raw output of a certain personality inventory or its normalized form. The second area for improvement would involve studying the use of non-western personality trait theories or indigenous measures. Such works are normally gauged towards understand a particular culture and would be a good area to apply linguistic analysis. Findings would be beneficial to both culture-specific and cross-cultural psychology.

References

[Coltheart, 1981] Max Coltheart. The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981.

[Friedman and Schustack, 2014] Howard S. Friedman and Miriam W. Schustack. *Personality: Classic Theories and Modern Research: Pearson New International Edition*. Pearson Education Limited, 2014.

[Golbeck *et al.*, 2011] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting Personality from Twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE, 2011.

[Goldberg, 1981] Lewis R. Goldberg. Language and Individual Differences: The Search for Universals in Personality Lexicons. *Review of Personality and Social Psychology*, 2(1):141–165, 1981.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[Havasi *et al.*, 2007] Catherine Havasi, Robert Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer, 2007.

[John *et al.*, 1991] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. The Big Five Inventory–Versions 4a and 54. Berkeley, CA: Univeristy of California, Berkeley, Institute of Personality and Social Research, 1991.

[John *et al.*, 2008] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. Paradigm Shift to the Integrative Big Five Trait Taxonomy. *Handbook of Personality: Theory and Research*, 3:114–158, 2008.

[Komarraju *et al.*, 2009] Meera Komarraju, Steven J. Karau, and Ronald R Schmeck. Role of the Big Five Personality Traits in Predicting College Students' Academic Motivation and Achievement. *Learning and Individual Differences*, 19(1):47–52, 2009.

[Landwehr *et al.*, 2005] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Machine Learning*, 59(1-2):161–205, 2005.

[Larsen and Buss, 2008] Randy J. Larsen and David M. Buss. *Personality Psychology: Domains of Knowledge About Human Nature*. McGraw Hill, 2008.

[Mairesse *et al.*, 2007] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, pages 457–500, 2007.

[McCrae and John, 1992] Robert R. McCrae and Oliver P. John. An Introduction to the Five-Factor Model and its Applications. *Journal of Personality*, 60(2):175–215, 1992.

[Mitchell, 1997] Tom M. Mitchell. *Machine learning. WCB*. McGraw-Hill Boston, MA:, 1997.

[Mohammad and Kiritchenko, 2013] Saif M. Mohammad and Svetlana Kiritchenko. Using Nuances of Emotion to Identify Personality. *arXiv preprint arXiv:1309.6352*, 2013.

[Mohammad and Turney, 2010] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[Norman, 1963] Warren T. Norman. Toward an Adequate Taxonomy of Personality Attributes: Replicated Factor Structure in Peer Nomination Personality Ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574, 1963.

[Park *et al.*, 2014] Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E.P. Seligman. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934, 2014.

[Peng *et al.*, 2015] Kuei-Hsiang Peng, Li-Heng Liou, Cheng-Shang Chang, and Duan-Shin Lee. Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts. In *Wireless and Optical Communication Conference (WOCC), 2015 24th*, pages 9–14. IEEE, 2015.

[Pennebaker and King, 1999] James W. Pennebaker and Laura A. King. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, 77(6):12961312, 1999.

[Pennebaker *et al.*, 2001] James W. Pennebaker, Martha E. Francis, and Roger J Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[Pennebaker *et al.*, 2003] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1):547–577, 2003.

[Pennebaker *et al.*, 2007] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. Operators Manual: Linguistic Inquiry and Word Count: LIWC2007. *Austin, Texas: LIWC.*, 2007.

[Poria *et al.*, 2013a] Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, and Newton Howard. Common Sense Knowledge Based Personality Recognition from Text. In *Advances in Soft Computing and Its Applications*, pages 484–496. Springer, 2013.

[Poria *et al.*, 2013b] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, (2):31–38, 2013.

[Richardson *et al.*, 2009] John D. Richardson, John W. Lounsbury, Tripti Bhaskar, Lucy W. Gibson, and Adam W. Drost. Personality Traits and Career Satisfaction of Health Care Professionals. *The Health Care Manager*, 28(3):218–226, 2009.

[Schwartz *et al.*, 2013] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[Smith, 2002] Lindsay I. Smith. A Tutorial on Principal Components Analysis. *Cornell University, USA*, 51(52):65, 2002.

[Turney and Littman, 2003] Peter D Turney and Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[Wilson, 1988] Michael Wilson. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10, 1988.